PATENT
200/1004-40

APPLICATION FOR UNITED STATES LETTERS PATENT

FOR

METHODS FOR ENRICHING POPULATIONS OF NUCLEIC ACID SAMPLES

Inventors:

Nila Patil, a citizen of the United States of America, residing in Woodside. California, USA

David Cox, a citizen of the United States of America, residing in Belmont. California, USA.

Charit Pethiyagoda, a citizen of Sri Lanka, residing in Sunnyvale. California, USA.

Andrew Sparks, a citizen of the United States of America, residing in Saratoga, California, USA.

Huang-Tsu Chen, a citizen of Taiwan, Republic of China, residing in Cupertino. USA.

Assignee:

Perlegen Sciences, Inc.
2021 Stierlin Ct.
Mountain View, CA 94043
650 625 4000
650 625 4574 (fax)

New Utility Patent Application

I hereby certify that this correspondence is being deposited with the U.S. Postal Service as Express Mail, Airbill #EV 322047524 US in an envelope addressed to: Mail Stop Patent Application, Commissioner for Patents, P.O. Box 1450, Alexandria, VA 22313-1450, on the date shown below.

Dated: 8/6/03 Signature: Paulette A. Ishii

METHODS FOR REDUCING COMPLEXITY OF NUCLEIC ACID SAMPLES

CROSS-REFERENCE TO RELATED APPLICATIONS

[0001] The present application derives priority from USSNs 60/228,251, filed August 26, 2000: 09/768,936 filed January 23, 2001: 09/938,878, filed August 24, 2001; and 10/131,832, filed April 24, 2002, which are incorporated by reference in their entirety for all purposes.

BACKGROUND

[0002] The scientific literature provides considerable discussion of nucleic acid probe arrays and their use in various forms of genetic analysis (see US Pat. Nos. 5,143,854, 5,252,743, 5,384,261, 5,405,783, 5,424,186, 5,445,943, 5,510,270, 5,677,195, 5,571,639, 5,837,832, 6,040,138, and 6,300,063 all incorporated herein by reference for all purposes). For example, nucleic acid probe arrays have been used for detecting variations in DNA sequences such as polymorphisms or species variations. Nucleic acid probe arrays have also been used for monitoring relative levels of populations of mRNA and detecting differentially expressed mRNAs.

[0003] Some methods for detecting polymorphisms using arrays of nucleic acid probes are described in WO 95/11995 (incorporated by reference in its entirety for all purposes), and a further strategy for detecting a polymorphism using an array of probes is described in EP 717,113. In this strategy, an array contains overlapping probes spanning a region of interest in a reference sequence. The array is hybridized to a labelled target sequence. Additional methods of polymorphism discovery and analysis are described in EP 0950720, which discusses use of primary arrays for de novo discovery of polymorphisms and use of secondary arrays for polymorphic profiling at the newly discovered polymorphic sites of different individuals. WO98/56954 discusses methods of identifying polymorphisms affecting expression of mRNA species.

[0004] Methods for using arrays of probes for monitoring expression of mRNA populations are described in US 6,040,138. Such methods employ groups of probes complementary to mRNA target sequences of interest. mRNA populations or amplification

products thereof are applied to an array, and targets of interest are identified, and optionally, quantified by determining the extent of specific binding to complementary probes. Additionally, binding of the target to probes known to be mismatched with the target can be used as a measure of background nonspecific binding and subtracted from specific binding of target to complementary probes. USSN 09/853,113, incorporated by reference for all purposes, discusses methods for determining functional regions in a genome using nucleic acid probe arrays.

[0005] However, the clarity and quality of the results obtained when using microarrays for analysis is, to a large degree, dependent on the quality and complexity of the target nucleic acid interrogated. The present invention provides methods for improving the quality and reducing the complexity of target nucleic acids applied to arrays, thereby improving the quality of the resulting data.

SUMMARY OF THE INVENTION

[0006] The present invention provides several methods for reducing the complexity of a population of nucleic acids prior to analyzing the nucleic acids on a microarray. Such reduction in complexity results in a subset of an initial population of nucleic acids where the subset is enriched for a desired property or lacks an undesired property. The resulting nucleic acids in the subset are then used as target DNA to be applied to a nucleic acid microarray for various types of analyses. Results obtained using a target sample of reduced complexity can be superior to those obtained using target samples where the methods of the present invention have not been employed. In general, the signal to noise ratio for samples with less complexity is much improved over untreated samples. The methods are particularly useful for analyzing nucleic acid populations having a high degree of complexity, for example, populations of DNA spanning a chromosome, DNA spanning a whole genome, or mRNA collections. Further, the methods of the present invention improve results obtained when pooling of target samples for analysis on an array. Pooling samples in appropriate circumstances leads to a reduction in cost and time of analysis if many samples must be analyzed.

[0007] Thus, one aspect of the present invention provides a method for analyzing a subset of nucleic acids within a nucleic acid population, comprising providing a population of nucleic acid fragments where at least some of these fragments have sequences that are

repeated. The population of nucleic acid fragments is denatured and incubated under conditions suitable to allow annealing of complementary sequences. The result after annealing is a mixture of double-stranded nucleic acids and single-stranded nucleic acids. Under annealing conditions. nucleic acid fragments containing repeat sequences preferentially anneal with one another relative to nucleic acid fragments lacking repeat sequences. Once annealing has taken place, the single-stranded nucleic acid fragments are separated from the double-stranded nucleic acid fragments, and the single-stranded nucleic acid fragments are then used as target DNA to be hybridized with probes on a nucleic acid probe array.

[0008] In another aspect of the invention there is provided a method for analyzing a subset of nucleic acids within a nucleic acid population, comprising providing a driver population of nucleic acids and a tester population of nucleic acids. The driver and tester populations are combined. denatured, and annealed. The result, as above, is a single-stranded subset of nucleic acids and a double-stranded subset of nucleic acids. Next, the driver set of nucleic acids in the mix are immobilized, resulting in unimmobilized single-stranded tester nucleic acids. immobilized double-stranded tester-driver or driver-driver nucleic acids and immobilized single-stranded driver nucleic acids. The unimmobilized single-stranded tester nucleic acids are separated from the immobilized nucleic acids and used as target DNA to be hybridized to probes on a nucleic acid probe array.

[0009] In yet another aspect of the present invention. there is provided a method of analyzing a subset of nucleic acids within a nucleic acid population, comprising providing a single-stranded driver population of nucleic acids and a single-stranded tester population of nucleic acids. The driver and tester nucleic acids are annealed. and the driver population is immobilized. The unimmobilized nucleic acids (primarily unhybridized tester nucleic acids) are separated from the immobilized nucleic acids (driver nucleic acids and tester nucleic acids complementary thereto). Once the unhybridized nucleic acids have been separated from the immobilized nucleic acids, the nucleic acids hybridized to the immobilized nucleic acids are dissociated and separated from the immobilized nucleic acids. These nucleic acids (primarily tester nucleic acids that are complementary to the driver nucleic acids) are then hybridized to probes on a microarray. This particular embodiment of the invention may be used in genotyping studies.

[0010] In yet another embodiment of the present invention, there are provided methods of using the polymerase chain reaction (PCR) to analyze a subset of nucleic acids

within a nucleic acid population. One such method comprises digestion of nucleic acid sample with a type IIs restriction endonuclease (or a combination of two or more type IIs restriction enzymes) to create fragments with single-stranded overhangs of a desired length and varied sequence, ligation of linkers that are specific to a subset of the overhangs produced by the restriction enzyme(s), and amplification of the selected nucleic acid fragments by PCR with linker-specific primer. The linkers used in this method contain the same PCR binding site, and only one PCR primer is required to amplify all fragments that are bound by a linker at each end. However, each ligation reaction must be separately carried out to select a particular subset of the original pool of nucleic acid fragments. Alternatively, using linkers that have different PCR primer binding site for each different overhang allows a single ligation reaction to be performed to anneal multiple different linkers to the pool of nucleic acid fragments. Selection of different subsets of the nucleic acid fragments is achieved by using different PCR primer pairs in subsequent PCR amplifications. Another such method is provided that does not require a restriction enzyme or a ligation reaction, comprising denaturation of the nucleic acid sample, annealing of a double-stranded branch primer to the denatured nucleic acid sample, extension of the primer by DNA polymerase, and a second round of denaturation, annealing, and primer extension, which is then followed by PCR or conditional PCR in which only fragment of certain size range will be amplified. Yet another such method comprises digesting a nucleic acid sample with a restriction enzyme that recognizes an interrupted palindromic sequence, annealing an adaptor containing fixed bases onto the ends of the resulting nucleic acid fragments, and then amplifying the adaptor-bound nucleic acid fragments using primers that contain fixed bases that are complementary to only a subset of the nucleic acid fragment. This particular embodiment of the invention may also be used in genotyping studies.

BRIEF DESCRIPTION OF THE FIGURES

[0011] Fig. 1 shows an exemplary scheme for removing repeat sequences from a population of nucleic acid fragments.

[0012] Fig. 2 shows an exemplary scheme for enriching a tester population of nucleic acids by hybridization of the tester population to a driver population of nucleic acids. In this scheme the driver DNA is a genomic clone in, for example, a BAC, YAC or PAC.

[0013] Fig. 3 shows an exemplary scheme for reducing complexity of genomic DNA in a tester population for further use in genotyping studies.

[0014] Fig. 4 shows an exemplary scheme for using PCR to reduce complexity of genomic DNA in a tester population for further use in genotyping studies. In scheme 4A-B, the linkers are designed to comprise the same PCR primer binding site. Alternatively, in scheme 4A-C the linkers are designed to comprise a different PCR primer binding site for each different overhang.

[0015] Fig. 5 shows another exemplary scheme for using PCR to reduce complexity of genomic DNA in a tester population for further use in genotyping studies. No restriction enzyme or ligation reaction is required in this scheme.

[0016] Fig. 6 shows yet another exemplary scheme for using PCR to reduce complexity of genomic DNA in a tester population for further use in genotyping studies. In this scheme, restriction enzyme(s) that recognizes an interrupted palindrome is used to digest the original nucleic acid sample.

DETAILED DESCRIPTION

[0017] Unless otherwise apparent from the context, reference to mRNA populations includes nucleic acid populations derived therefrom by processes in which the mRNA serves as template for polynucleotide extension, such as cDNA or cRNA.

[0018] A nucleic acid is a deoxyribonucleotide or ribonucleotide polymer in either single- or double-stranded form, including known analogs of natural nucleotides unless otherwise indicated.

[0019] An oligonucleotide is a single-stranded nucleic acid ranging in length from 2 to about 500 bases.

[0020] A probe is a nucleic acid capable of binding to a target nucleic acid of complementary sequence through one or more types of chemical bonds, usually through complementary base pairing, usually through hydrogen bond formation. A nucleic acid probe may include natural (i.e. A, G, C, or T) or modified bases (e.g., 7-deazaguanosine, inosine). In addition, the bases in a nucleic acid probe may be joined by a linkage other than a phosphodiester bond, so long as it does not interfere with hybridization. Thus, nucleic acid probes may be peptide nucleic acids in which the constituent bases are joined by peptide bonds rather than phosphodiester linkages.

[0021] Specific hybridization refers to the binding, duplexing, or hybridizing of a molecule preferentially to a particular nucleotide sequence when that sequence is present in a complex mixture (e.g., total cellular) DNA or RNA. Stringent conditions are sequence-dependent and are different in different circumstances. Generally, stringent conditions are selected to be about 5°C lower than the thermal melting point (Tm) for the specific sequence at a defined ionic strength and pH. The Tm is the temperature (under defined ionic strength, pH, and nucleic acid concentration) at which 50% of the probes complementary to the target sequence hybridize to the target sequence at equilibrium. As the target sequences are generally present in excess, at Tm, 50% of the probes are occupied at equilibrium. Typically, stringent conditions include a salt concentration of at least about 0.01 to 1.0 M Na ion concentration (or other salts) at pH 7.0 to 8.3 and the temperature is at least about 30°C for short probes (e.g., 10 to 50 nucleotides). Stringent conditions can also be achieved with the addition of destabilizing agents such as formamide. For example, conditions of 5X SSPE (750 mM NaCl, 50 mM NaPhosphate, 5 mM EDTA, pH 7.4) and a temperature of 25-30 °C are suitable for allele-specific probe hybridizations.

[0022] A perfectly matched probe has a sequence perfectly complementary to a particular target sequence. A test probe is typically perfectly complementary to a portion (subsequence) of the target sequence. The term "mismatch probe" refers to probes whose sequence is deliberately selected not to be perfectly complementary to a particular target sequence. Although the mismatch(es) may be located anywhere in the mismatch probe, terminal mismatches are less desirable as a terminal mismatch is less likely to prevent hybridization of the target sequence. Thus, probes are often designed to have the mismatch located at or near the center of the probe such that the mismatch is most likely to destabilize the duplex with the target sequence under the test hybridization conditions.

[0023] A polymorphic marker or site is the locus at which divergence occurs. Preferred markers have at least two alleles, each occurring at frequency of greater than 1%, and more preferably greater than 10% or 20% of a selected population. A polymorphic locus may be as small as one base pair. Polymorphic markers include restriction fragment length polymorphisms, variable number of tandem repeats (VNTR's), hypervariable regions, minisatellites, dinucleotide repeats, trinucleotide repeats, tetranucleotide repeats, simple sequence repeats, and insertion elements such as Alu. The first identified allelic form is arbitrarily designated as the reference form and other allelic forms are designated as alternative or variant alleles. The allelic form occurring most frequently in a selected population is sometimes referred to as the wildtype form. Diploid organisms may be homozygous or heterozygous for allelic forms. A diallelic polymorphism has two forms. A triallelic polymorphism has three forms. A single nucleotide polymorphism (SNP) occurs at a polymorphic site occupied by a single nucleotide, which is the site of variation between allelic sequences. The site is usually preceded by and followed by highly conserved sequences of the allele (e.g., sequences that vary in less than 1/100 or 1/1000 members of the populations). A single nucleotide polymorphism usually arises due to substitution of one nucleotide for another at the polymorphic site. Single nucleotide polymorphisms can also arise from a deletion of a nucleotide or an insertion of a nucleotide relative to a reference allele.

[0024] The present invention provides several methods for reducing the complexity of a population of nucleic acids prior to performing an analysis of the nucleic acids on a nucleic acid probe array. The results obtained using nucleic acid array technologies are enhanced by reducing complexity of the target or sample nucleic acids applied to the array. The methods result in a subset of the initial population enriched for a desired property, or lacking nucleic acids having an undesired property, and the resulting nucleic acids in the subset are then applied to the array for various types of analyses. The methods are particularly useful using nucleic acid probe arrays to analyze nucleic acid populations having a high degree of complexity, for example, populations of chromosomal DNA, or whole genomic DNA, or mRNA. The methods of the present invention attain reduced complexity of samples which enables analysis of pooled samples.

[0025] In some methods, an initial population of nucleic acids is treated so as to reduce or eliminate fragments having repeat sequences. In general, nonrepeat sequences contain the coding and key regulatory regions of genomic DNA and are of interest for most

subsequent genetic analyses. Repeat sequences can be eliminated by a process that involves denaturing the initial population of nucleic acids. if double-stranded, and reannealing. Single stranded nucleic acids with repeat sequences preferentially hybridize with each other relative to single stranded nucleic acids of unique sequence because there is a greater probability of nucleic acids with repeated regions finding a complementary nucleic acid with which to hybridize.

[0026] After annealing, double-stranded (annealed) and single-stranded nucleic acids are separated from one another. The resulting separated single-stranded nucleic acids are enriched for nonrepeat sequences. These enriched single-stranded sequences are then applied to a nucleic acid microarray for a variety of genetic analyses. For example, such analyses include de novo polymorphic site discovery. detection of a plurality of predetermined polymorphic sites. SNP analysis, expression analysis and the like. In general, when analyzing arrays. it is desirable to discriminate between specific hybridization between the microarray probes and target sequences and nonspecific hybridization between the probes and target sequences. Reducing the complexity of the target nucleic acid leads to reduction in non-specific hybridization, resulting in less "noise" or background. Increasing the signal to noise ratio is an extremely important factor in microarray—particularly when analyzing target samples that may have low copy numbers of some sequences.

[0027] Repeat sequences are sequences that occur more than once in a haploid genome of a single organism. In some instances. multiple copies of a repeat sequence are identical. In other instances, there are some divergences between copies but substantial sequence identity. e.g., at least 80 or 90%. More than 30% of human DNA consists of sequences repeated at least 20 times. Families of repeated DNA sequences of 100-500 bp that are interspersed throughout the genome are sometimes known as SINES (short interspersed repeats). Alu sequences are examples of SINES that are about 300 bp and occur almost 1 million times in the human genome. Longer interspersed repeat sequences of 1 kb or more are known as LINES (long interspersed repeats). Some repeat sequences are not interspersed throughout the genome but are concentrated at particular loci. These repeats are known as satellite repeats. Some repeat sequences are actual genes, such as the genes that code for ribosomal RNAs and histones. However, the function, if any, of most repeat sequences is unclear. The vast majority of protein coding sequences and their associated regulatory sequences occur in single copy regions of the genome.

[0028] Thus, one aspect of the present invention provides methods for enriching for single copy regions of a genome relative to repeat sequences before performing a genetic analysis using a nucleic acid probe array. Figure 1 shows a schematic of this aspect of the invention. First, a population of genomic DNA (101) is fragmented (102) by digestion with a restriction enzyme or DNaseI to produce fragments of, for example, an average size of about 300 bp (103). The fragments are denatured and allowed to reanneal (104). Repeat sequences hybridize with each other, whereas nonrepeat sequences remain in single stranded form (105). The double-stranded hybrids and the single-stranded sequences are then separated on a hydroxyapatite HPLC column (106). The DNA is loaded in a phosphate buffer and eluted using a phosphate buffer gradient. As seen in 107, single-stranded DNA elutes at a concentration of about 120-140 mM phosphate, and double-stranded DNA elutes at a concentration of about 500mM to 1 M phosphate. The single-stranded sequences may be labeled then applied to a microarray.

[0029] The starting population of nucleic acids for enrichment (101) can be from a genomic DNA from a whole genome, a collection of chromosomes, a single chromosome, or one or more regions from one or more chromosomes, or cloned DNA, RNA or cDNA. Genomic DNA can be obtained from virtually any tissue source (other than pure red blood cells). For example, convenient tissue samples include whole blood, semen, saliva, tears, urine, fecal material, sweat, buccal, skin and hair. The nucleic acids may be obtained from the same individual, which can be a human or other mammal or other species, or from different individuals of the same species or different individuals of different species.

[0030] Both enzymatic and mechanical methods can be used for fragmentation (102). The fragmenting can be effected by restriction digestion, often using a partial digest with a restriction enzyme with a short recognition site or a limited digest with a mixture of enzymes or with DNaseI. Alternatively, fragments can be produced by sonication, or by PCR amplification using random primers or random fragments of an initial substrate. Other suitable methods include mechanic or liquid shearing by using a French press or a UCHGR Shearing Device. In some methods, the fragments are overlapping fragments spanning a length of 100 kb, 1 Mb, 10 Mb or 100 Mb. Also, the initial substrate can be amplified, and/or labeled before or after fragmentation. In some methods, fragments are attached to linkers at one or both ends to provide primer sites for subsequent amplification. Fragments may have an average size of about 300 bp. For example, appropriate restriction enzymes may be used to cut genomic DNAs to a desired range of sizes.

[0031] Fragments containing repeat sequences are removed from the population by a combination of denaturation (assuming the fragments are double stranded) and reannealing (104). Denaturation can be effected by heating fragments in excess of the average melting point of the fragments. The denatured fragments are then cooled to below the average melting point (e.g., about 25 degrees below the average melting point) for reannealing. The reassociation can be followed by, for example, monitoring hyperchromicity at 260 nm. As DNA renatures, the hyperchromicity increases due to greater absorbance of double-stranded DNA relative to single-stranded DNA. The hyperchromicity curve shows a point of inflexion at which half of the DNA is reannealed. The reannealing reaction is often stopped about this time, but the duration of the reaction can be adjusted depending on the percentage of repetitive DNA in the sample. For example, the more repetitive DNA sequences present in a sample, the longer the annealing reaction should proceed. The reannealing reaction effectively can be stopped by rapid cooling of the annealing mixture to just above freezing.

[0032] After the annealing reaction (105), annealed double-stranded DNA is separated from single-stranded DNA (106). Separation can be effected using column chromatography. A hydroxyapatite (calcium phosphate) column is particularly suitable (see Sambrook, *et al.*, Molecular Cloning: A Laboratory Manual, A8.32 (Cold Spring Harbor Laboratory, New York) (1989)). Both single- and double- stranded nucleic acids bind to the column at low phosphate concentration (10-30 mM sodium phosphate). At intermediate phosphate concentrations (120 mM to 140 mM,), single-stranded DNA no longer binds the column, however, double-stranded DNA continues to bind. At higher concentrations (400 mM), both single- and double-stranded DNA no longer bind to the column. Thus, DNA can be loaded on the column at low phosphate concentration, in which case both single- and double-stranded nucleic acids bind. Single-stranded nucleic acids are then eluted with an increasing concentration gradient of sodium phosphate buffer. Alternatively, single- and double-stranded nucleic acids can be loaded at an intermediate phosphate concentration, in which case the single-stranded nucleic acids pass though without binding and the double-stranded nucleic acids bind (see Genome Analysis: A Laboratory Manual; Volume 2, Detecting Genes (eds. Bruce Birren et al., Cold Spring Harbor Press, 1998) and Sambrook, *et al.*, Molecular Cloning: A Laboratory Manual (Cold Spring Harbor Laboratory, New York) (1989)). In some methods, hydroxyapatite columns are combined with HPLC. Alternatively or additionally, the annealing reaction mixture can be treated with a nuclease that selectively digests double-stranded DNA.

[0033] After separation of single-stranded nucleic acids from double-stranded nucleic acids (107), the single-stranded nucleic acids can be applied directly to a microarray, or can be the subject of additional treatment (for example, labeling reactions or amplification reactions) before application to the array. For example, in some methods, the single-stranded fragments are allowed to anneal with each other, forming double-stranded fragments, which are then amplified, labelled. and denatured before being applied to the microarray. In some methods, single-stranded nucleic acids that were not previously labeled are now labelled before application to the microarray. Some methods for end-labelling fragments are described by WO97/27317. In some methods, the single-stranded fragments are broken down to still smaller fragments before being applied to an array.

[0034] The type of array to which the fragments are applied of course depends on the form of contemplated analysis. In some methods, fragments are applied to arrays designed for de novo polymorphism discovery. These arrays typically contain overlapping probes tiling a region of a known reference sequence. The hybridization pattern of the fragments to the array indicates the site and nature of points of divergence between the sequence of the fragments and the reference sequence, and hence the location and identity of polymorphic sites. In other methods. the fragments are applied to an array designed to detect a collection of polymorphisms where the location and nature of polymorphic forms is already known. In such methods, the hybridization pattern of the nucleic acid fragments to the array indicates a polymorphic profile of the individual from whom the fragments were obtained (i.e., a matrix of polymorphic sites, and polymorphic forms present in those sites). Microarray fabrication, design and the uses thereof are disclosed in US Pat. Nos. 5,143,854, 5,252.743. 5,384.261, 5,405,783, 5,424,186, 5.445,943, 5,510,270, 5,677,195, 5,571,639, 5,837,832. 6.040,138, and 6,300,063 all incorporated herein by reference for all purposes.

[0035] As mentioned previously, the nucleic acid samples can be amplified before or after enrichment. For example, an individual genomic DNA segment from the same genomic location as a designated reference sequence can be amplified by using primers flanking the reference sequence. Multiple genomic segments corresponding to multiple reference sequences can be prepared by multiplex amplification including primer pairs flanking each reference sequence in the amplification mix. Alternatively, the entire genome can be amplified using random primers (typically hexamers) (see Barrett et al., *NAR* 23:3488-3492 (1995)) or by fragmentation and reassembly (see, e.g., Stemmer et al., *Gene* 164:49-53 (1995)). RNA samples are also often subject to amplification. In this case amplification is

typically preceded by reverse transcription. Amplification of all expressed mRNA can be performed, for example, as described by WO 96/14839 and WO 97/01603.

[0036] The PCR method of amplification is well known in the art and described in PCR Technology: Principles and Applications for DNA Amplification (ed. H.A. Erlich, Freeman Press, NY, NY, 1992); PCR Protocols: A Guide to Methods and Applications (eds. Innis, et al., Academic Press, San Diego, CA, 1990) and U.S. Patent 4,683,202, each of which is incorporated by reference for all purposes. Further, nucleic acids in a target sample can be labeled in the course of amplification by inclusion of one or more labeled nucleotides in the amplification mix. Alternatively, labels can be attached to amplification products after amplification, for example, by end-labeling. The amplification product can be RNA or DNA depending on the enzyme and substrates used in the amplification reaction.

[0037] Other suitable amplification methods include the ligase chain reaction (LCR) (see Wu and Wallace, *Genomics* 4:560 (1989), Landegren et al., *Science* 241:1077 (1988)), transcription amplification (Kwoh et al., *PNAS USA* 86:1173 (1989)), self-sustained sequence replication (Guatelli et al., *PNAS USA*, 87:1874 (1990)) and nucleic acid based sequence amplification (NASBA). The latter two amplification methods involve isothermal reactions based on isothermal transcription, which produce amplification products of both single stranded RNA (ssRNA) and double stranded DNA (dsDNA) in a ratio of about 30 or 100 to 1, respectively.

[0038] Figure 1 illustrates separation of repeat sequences from other sequences in a nucleic acid from a single source. In addition, a variety of enrichments can be performed by hybridization of a nucleic acid sample from one source to a nucleic acid sample from a different source. Herein, a nucleic acid sample from one source or sources may be referred to as a "tester nucleic acid" and a nucleic acid sample from another source or sources is referred to as a "driver nucleic acid". An example of using tester and driver nucleic acids to reduce complexity of a nucleic acid sample is shown in Fig. 2. In this example, the driver nucleic acids (201) are genomic DNA, genomic clones, BACs, YACs or PACs, and the tester nucleic acid is RNA (202). In step 203, the RNA is subjected to reverse transcription to produce cDNA (204). In step 205, the driver nucleic acids are cleaved using a restriction enzyme and ligated to linkers containing primer sites to produce fragments of average size about 300 bp (207). Similarly, in step 206, the tester nucleic acids (represented now by cDNA) are cleaved using a restriction enzyme and ligated to linkers containing primer sites to produce fragments of average size about 300 bp (208).

[0039] The driver nucleic acid fragments (207) are then amplified in the presence of biotin labeled nucleotides (209) to produce biotin labeled fragments (210) (only one strand of the amplified fragments of the driver nucleic acids is shown in 210). The tester nucleic acid fragments (fragmented cDNA) (208) are amplified (211) to produce amplified tester fragments (212). The biotin-labeled driver fragments (210) are combined with the amplified cDNA fragments (212) and denatured then allowed to hybridize to each other in solution (213). The biotin-labeled driver fragments and any hybridized tester fragments are then immobilized to streptavidin labeled magnetic beads by virtue of the affinity of the streptavidin for the biotin label on the driver nucleic acids. The bead/hybrid complexes are then washed to remove unhybridized tester nucleic acids. For some purposes, such as looking for unique tester sequences, the unhybridized tester nucleic acids may be of interest and are retained for analysis. For other purposes, such as examining sequences common to both tester and driver nucleic acids, the hybridized tester nucleic acids are dissociated from the bead/immobilized driver complex (217) and the eluted tester nucleic acids (218) are analyzed. In general in these methods, either or both driver and tester nucleic acids can be amplified before the enrichment procedure.

[0040] Fragmentation (steps 205 and 206) can be achieved by any of the methods described above, usually to an average size of about 200-700 bp or about 250-500 bp. Fragmentation before enrichment is typical with genomic populations and possible, but not usual, with mRNA populations. In some embodiments, a population of nucleic acids is fragmented, the fragments are ligated to oligonucleotides having primer sites, and the ligated fragments are amplified. Also, in alternative embodiments of the present invention, the tester nucleic acid fragments can be labeled instead of the driver fragments or in addition to the driver fragments. Alternatively, labeling can be performed before or after the enrichment procedure.

[0041] Also in these methods, populations of driver and tester nucleic acid fragments are denatured if initially double-stranded. Denaturation can take place before or after combining the two fragment populations. If the populations of fragments are labeled separately, generally they are mixed after denaturation and allowed to reanneal. As in the methods for eliminating repeat sequences within a single nucleic acid population, denaturation can be performed by raising the temperature over the average melting point of driver and tester nucleic acid populations.

[0042] Hybrids between tester and driver nucleic acids are separated from unhybridized tester nucleic acid. As shown in Figure 2, separation can be effected by inclusion of a tag on all driver fragments and immobilizing the driver fragments to a binding moiety. For example, a biotin tag can be attached to driver fragments by amplifying them using a biotin labelled primer or biotin labelled nucleotides or by ligating them to biotin labeled oligonucleotides or by directly attaching biotin to the fragments. Biotin labeled driver fragments can then be immobilized to a support bearing an avidin or streptavidin binding moiety. For example, magnetic beads coated with streptavidin, available from Dynal (Norway), are suitable for immobilizing biotin-labelled DNA. Procedures for performing enrichments of cDNA using immobilized DNA on beads are described by Birren et al., *supra* at ch. 3. Other combinations of tag and binding moiety similarly can be used. Alternatively, hybrids can be separated from single-stranded fragments using hydroxyapatite chromatography as described above. As yet another alternative, separation can be effected using a nuclease that digests duplex nucleic acids without digesting single stranded nucleic acids or vice versa. For example, S1 nuclease preferentially digests single stranded DNA, whereas most restriction enzymes preferentially digest double stranded DNA.

Driver: Genomic DNA/Tester:mRNA

[0043] In some methods, the driver population is genomic DNA and the tester population is an mRNA population or nucleic acid population derived therefrom (e.g., cDNA or cRNA). As will become apparent, such methods serve to normalize the representation of different nucleic acid sequence species within the mRNA population (or nucleic acids derived therefrom). In other words, the methods enrich the representation of rare mRNA species relative to the more common mRNA species. In such methods, the driver population can be from a whole genome, a chromosome, a collection of chromosomes or one or more regions of one or more chromosomes. If an entire genome is included, then the resulting enriched population of mRNAs includes mRNAs spread throughout the genome. If a single chromosome is included, then the enriched population of mRNAs is restricted to mRNAs hybridizing to that chromosome, and so forth. The mRNA population used as the tester population can be from a single tissue type, from a cell line or from a mixture of tissue types. If from a single tissue type, the mRNA population and the resulting enriched population contains a bias toward the mRNAs expressed in that cell type. If the mRNA population is from a representative mixture of tissue types, then the population and the subsequent enriched

populations contains most or substantially all (e.g., at least 50% , 75% or 90%) of mRNAs expressed by the organism. Some cell lines, such as HeLa cells, also express a substantial proportion of all mRNAs typically expressed in an organism. If cDNA or cRNA is prepared from mRNA, the preparation can be performed under conditions that preserve the relative representations of mRNA species in the original population as described by USSN 6,040,138. However, such is generally not necessary because the proportions are, of course, deliberately changed in the enrichment procedure. Thus, conventional methods of cDNA preparation using polyT primers or random hexamers can be used (see Birren et al., *supra* at ch. 3). In some methods, adapters are ligated to cDNA to facilitate subsequent amplification or labelling.

[0044] When driver genomic DNA is hybridized with tester mRNA (or a nucleic acid derived therefrom), the mRNA hybridizes to complementary sequences in the genomic DNA sequences. However, in general, each mRNA species has only a single complementary genomic DNA sequence in a haploid genome. Accordingly, highly represented mRNA species and minimally represented species (and intermediately represented sequences) in general all hybridize to genomic DNA to a similar extent. In theory, one molecule of mRNA should hybridize per haploid genome for a single copy gene. In practice, this ratio is not observed for all single copy genes due to the presence of introns. For example, a gene having ten spaced exons can hybridize to different regions of ten copies of the same mRNA. Nevertheless, the hybridization does result in substantial normalization between mRNA species. For example, whereas the variation copy number between species in an unnormalized population can be greater than 10^5, in a normalized population, the variation is more typically within a factor of 1000, 100, or 10.

[0045] After performing hybridization, the double-stranded nucleic acids are separated from single-stranded nucleic acids. If the driver nucleic acid population is labeled with a binding moiety used for separation, then all driver nucleic acids are captured, including driver/tester hybrids, and the only unbound single-stranded nucleic acids are unhybridized tester. The unhybridized tester is set aside. Then the tester nucleic acids that hybridized to the driver nucleic acids are dissociated from the complementary driver nucleic acids (e.g., by raising the temperature above the melting point). The driver nucleic acids remain bound (generally associated with the solid phase) and the subset of complementary tester nucleic acids is obtained in solution in single-stranded form. The single-stranded fragments can be labeled (if not labeled already) and applied directly to an array.

Alternatively, the fragments can be renatured with each other. for amplification and labeling. Amplified fragments are then denatured again before being applied to an array.

[0046] The subset of tester fragments obtained can be subject to a variety of genetic analyses. In some methods, the fragments are used for de novo polymorphism discovery in a fashion similar to that described above. The polymorphisms discovered thereby are highly likely to occur within expressed regions of the genome. The subset of tester fragments also can be used for polymorphic profiling of previously characterized polymorphic sites within expressed regions within an individual. Use of mRNA populations has advantages relative to use of genomic DNA in that nonexpressed regions of the genome. which probably contain relatively few polymorphic sites of functional significance but which would otherwise contribute to a background of nonspecific binding on the array, are not applied to the array. It is estimated that only 5% of the human genome contains coding regions.

[0047] The subset of tester fragments also can be used for discovering relatively rare differentially expressed genes. For example, by comparing tester populations enriched as described above from different tissue types, one can identify species within one tester mRNA population that are not expressed in another mRNA population. Such mRNA species can be cloned as described in Sambrook, *et al.*, Molecular Cloning: A Laboratory Manual (Cold Spring Harbor Laboratory, New York) (1989), incorporated herein by reference. This type of analysis is particularly useful for identifying genes that are expressed at low levels or not at all in a tissue.

Driver: Genomic DNA, PCR Product or Clone/Tester: Genomic DNA, PCR Product or Clone

[0048] In some methods, both driver and tester populations are genomic DNA but from different sources. In some methods, the different sources are different individuals from the same species, in others, the different sources are individuals from different species. For example, the two sources can be two different humans, or one human and one cat, or one mouse and one dog, and so forth. Such methods may serve to enrich either fragments that are common to the two sources or to enrich fragments that differ between the two sources. For the former type of enrichment, one retains tester fragments hybridizing to driver fragments. For the latter type of enrichment, one retains tester fragments not hybridizing to driver fragments. Common sequences are of interest because commonality often implies evolutionary conservation: hence, a possible important functional role. Polymorphisms

occurring within regions that are conserved between species are more likely to have phenotypic consequences. Accordingly, given the vast number of polymorphic sites within a genome, it can be advantageous to focus on conserved regions for polymorphism discovery, and/or to use polymorphisms within conserved regions for association studies. Disparate sequences between sources are also of interest, because these sequences are the locus of genetic diversity between different individuals and/or species.

[0049] In these methods, as in other methods, driver and tester populations can be obtained from whole genomes, collections of chromosomes, individual chromosomes or one or more regions of individual chromosomes. Usually, the fragments within a driver population are obtained from the same individual, as is the case for the fragments within a tester population; however, the driver and tester populations are generally obtained from different individuals. Either driver and/or tester populations can be amplified before performing hybridization. The tester population can be labelled before or after the hybridization. If the goal is to isolate sequences that are common between the driver and tester populations, the nonhybridizing subset of nucleic acids from the tester population are set aside, and the subset of tester fragments hybridizing to the driver are dissociated from the driver. These fragments can be subject to amplification and/or labelling before being applied to an array. If the goal is to isolate disparate fragments between the driver and tester populations, then the driver and tester fragments that hybridize are set aside and the nonhybridizing tester fragments are applied to an array (optionally with labelling, if not already labelled). Alternatively, the nonhybridizing tester fragments can be hybridized with each other, amplified and labeled before being applied to an array.

[0050] In other methods, hybridization between driver and tester fragments is used to selectively amplify regions of genomic DNA to be used in genotyping studies. The goal in such methods is to apply one or more regions of genomic DNA of interest to an array without applying regions that are not of interest. In other methods in the art, such a goal could be achieved by selective amplification of the desired genomic regions. However, performing selective amplification on a large number of samples from multiple noncontiguous regions can be tedious and subject to error. In the present method, the amplification can be performed on a single genomic sample, and the amplified sample then is used as a driver population to enrich equivalent regions from a broader population of tester genomic DNA. For example, the driver population can be a long-range or short-range PCR product of a particular chromosome or oligonucleotides with a sequence or sequences of choice. The

tester population can be a whole genomic population or the chromosomal region from which the long- or short-range PCR product was obtained. When the tester genomic DNA population is annealed with the driver PCR product population, substantially only fragments from the tester population that are complementary to the driver population will hybridize thereby reducing complexity of the tester population. These fragments can then be dissociated from the driver and applied to an array.

[0051] Figure 3 shows a process as just described where the driver nucleic acid is a PCR product and the tester nucleic acid is genomic DNA. In this example, the driver nucleic acids (301) are long- or short-range PCR products, and the tester nucleic acid is genomic DNA (302). In step 303, the PCR products are, optionally, fragmented, then tagged or labeled (304) with, e.g., biotin labeled nucleotides, to immobilize the driver DNA and facilitate separation of the driver DNA from the tester DNA in a later step. The genomic tester DNA can be DNA from a single individual or a pooled sample of DNA from many individuals. In step 305, the tester DNA is fragmented and, optionally, the fragments are blunt ended by incubation with dNTPs and T4 DNA polymerase or Klenow or the like to fill in the fragments' termini. After blunt-ending, linkers containing primer sites may be ligated to the tester fragments (306). At this point, the tester fragments may be subjected to an optional amplification step using the primer sites in the ligated linkers.

[0052] At step 310, the tagged driver PCR products and the tester fragments are combined, denatured and annealed to produce a mixture of annealed products (311). The biotin-labeled driver PCR products and any tester fragments hybridized to the driver PCR products are immobilized to streptavidin labeled magnetic beads (312) by virtue of the affinity of the streptavidin for the biotin label on the driver nucleic acids. The bead/hybrid complexes (313) are washed (314) to remove unhybridized tester nucleic acids and then the tester fragments that hybridized to the bead/hybrid complex are eluted (314). The resulting population of eluted tester fragments (315) contains genomic DNA sequences that are complementary or nearly complementary to the specific PCR sequences of the driver population. At this juncture, the tester genomic DNA may be subjected to amplification and then labeled, and the final tester product (317) is analyzed by hybridization to any array.

[0053] Fragmentation (optional step 303 and step 305) can be achieved by any of the methods described above and results in an average fragment size of about 50-700 bp or about 250-500 bp. Also in these methods, the populations of driver and tester nucleic acid fragments are denatured either before or after combining the two fragment populations. As in

the methods described previously, denaturation can be performed by raising the temperature over the average melting point of driver and tester nucleic acid populations. In the example in Figure 3, again the separation step is effected by inclusion of a biotin tag on all driver fragments and immobilizing the driver fragments with a streptavidin binding moiety coupled to a support. However, other combinations of tag and binding moiety can be used.

[0054] The tester fragments obtained can be used for polymorphic profiling. The benefits of such enrichment are particularly evident when it desired to analyze a plurality of noncontiguous regions within a genome (e.g., ten or more), and/or when it desired to analyze tester DNA from a plurality of individuals (e.g., ten or more). For example, the driver PCR products may be selected to be those from regions that contain one or more single nucleotide polymorphisms (SNPs). These SNP-containing driver molecules can be used as "hooks" to fish out the complementary SNP-containing regions of the genomic tester DNA. Hybridization conditions can be varied such that complementary tester sequences will hybridize to driver sequences even with one or more mismatches. In such a case, it would not be necessary for the driver DNA population to be heterozygous for each SNP. Thus, a driver PCR product population containing many SNPs is able to fish out many SNP-containing tester genomic DNA fragments regardless of the SNP allele present. Once the genomic tester SNP-containing fragments have been "hooked", they can be eluted, amplified, labeled (by nick translation or by end labeling), and applied to an array.

[0055] Alternatively, the driver nucleic acid can be genomic DNA with repeat regions and the tester DNA can be genomic DNA from a sample of interest. In this embodiment, the driver nucleic acid will "fish out" (hybridize with) the repetitive regions in the tester nucleic acid, and the tester genomic DNA that does not hybridize to the driver DNA (tester genomic DNA subtracted of repetitive sequences) would ultimately be applied to an array. Optionally, after being depleted of repetitive sequences, the subtracted tester nucleic acids are subjected to further processing to further reduce complexity before hybridization to an array. Such processing may, for example, include amplification of specific regions using random primers or sequence-specific primers. In the example in the preceding paragraph, the driver DNA is used as a hook to fish out the tester sequences that are to be used for analysis; however in this example, the driver DNA is used as a hook to fish out the tester sequences that are to be removed from analysis. In both cases, the complexity of the tester nucleic acids that are ultimately analyzed is reduced.

[0056] The present invention thus provides a method to reduce complexity of a genomic DNA sample for polymorphic profiling on arrays. Because the DNA applied to the array contains selected sequences. background is reduced significantly.

Driver: mRNA/Tester: Genomic DNA

[0057] In other methods, a driver population of mRNA or nucleic acids derived therefrom is used to enrich a tester population of genomic DNA. Such methods enrich the genomic DNA population for fragments represented in the mRNA. The enrichment results in a population of nucleic acids that are normalized in copy number relative to the original population of mRNA. In addition, the enriched nucleic acids include regions of genomic DNA proximate to expressed regions, such as intron-exon borders, and nonexpressed regulatory sequences, such as promoters and enhancers. The enriched population can be used in similar analyses to those described above. In addition. the population is useful for discovering and detecting polymorphisms in nonexpressed regions of DNA adjacent to the expressed regions that cannot be detected by analysis of mRNA populations. Such polymorphisms may have roles in regulating the extent of expression of a gene.

[0058] The tester population can be from a whole genome. a chromosome, a collection of chromosomes or one or more regions of one or more chromosomes. If an entire genome is included. then the enriched population of nucleic acids typically includes nucleic acids spread throughout the genome. If a single chromosome is included, then the enriched population of nucleic acids is. of course, within this chromosome. An mRNA population used as the driver population can be from a single tissue type, from a cell line or from a mixture of tissue types. also as described above. After hybridization of driver and tester populations, unhybridized tester fragments are set aside. Hybridized tester fragments are dissociated from the driver fragments. The resulting tester fragments can the be applied to an array (optionally with labelling. if not already labelled). Alternatively, the resulting tester fragments can be renatured. amplified, and optionally, labelled before being applied to an array.

Driver: mRNA/Tester: mRNA

[0059] In some methods, both driver and tester populations are mRNA populations from different sources. The different sources can be different tissues from an individual or individuals within the same species. Alternatively, the different sources can be the same

tissue type from different species, (e.g., human and mouse, cat, dog, horse. cow. sheep, primate and so forth). In a further variation, the two sources can be the same tissue subject to different environmental factors, for example, exposure to a drug or potentially toxic compound. The enrichment can be used to enrich either for fragments that are common to the two populations or for fragments that are differentially represented between the two populations. Fragments that are common to two populations of mRNA from the different species presumably are enriched for sequences that have been subject to evolutionary conservation. As previously discussed, polymorphisms within such sequences are particularly likely to have phenotypic consequences. Accordingly, such common fragments are useful for de novo polymorphism discovery and profiling of previously characterized polymorphisms.

[0060] Differentially expressed mRNA species can also be used for polymorphism analysis, or be applied to expression monitoring arrays for identification and further characterization of the genes encoding such mRNA species. For example, such mRNA species can be applied to probe arrays containing large numbers of random probes. Probes showing specific hybridization can then be used as primers or probes to isolate genes responsible for differentially expressed mRNAs. Alternatively, the mRNA species can be hybridized to an expression monitoring array containing probes for known mRNA species. If the mixture of differentially expressed mRNAs resulting from enrichment is one of the known mRNA species. this is indicated by the resulting hybridization pattern.

[0061] As in other methods, common mRNA species between the two populations are isolated by separating the nonhybridizing tester mRNA fragments from the hybridizing double-stranded fragments. dissociating the double-stranded fragments and separating the tester mRNA from driver mRNA. In addition, the dissociated tester mRNA can be subjected to amplification and labelling before application to an array. Amplification. if any, can be conducted with or without preservation of relative copy number of amplified species.

[0062] As previously discussed, a variety of probe array designs can be used in the invention depending on the intended type of genetic analysis. Probe arrays and their uses are disclosed in US Pat. Nos. 5,143,854, 5,252,743, 5,384.261, 5,405,783, 5,424,186, 5,445,943, 5,510,270, 5,677,195, 5,571,639, 5,837,832, 6,040.138. and 6,300,063, all incorporated herein by reference for all purposes. Some arrays are designed for de novo discovery of polymorphisms. Such arrays contain at least a first set of probes that covers or "tiles" one or more reference sequences (or regions of interest therein). and the reference sequence can be a

chromosome, a genome, or any part thereof. Tiling means that the probe set contains overlapping probes that are complementary to and span a region of interest in the reference sequence. For example, a probe set might contain a ladder of probes, each of which differs from its predecessor in the omission of a 5' base and the acquisition of an additional 3' base. The probes in a probe set may or may not be the same length. Such arrays typically contain at least one probe for each base to be analyzed.

PCR-based Methods of Complexity Reduction

[0063] Analysis of nucleic acid samples (e.g. human genomic DNA) often requires techniques to increase the relative concentration of a subset of the original nucleic acid sample for further analysis or manipulation (complexity reduction). This may be accomplished by either removing a subset that is not desired, or by directly increasing the concentration of the subset that is desired, for example, using the polymerase chain reaction (PCR). The need for complexity reduction is often due to inherent technical limitations of the analysis technology to be used. Technologies traditionally requiring complexity reduction of human genomic DNA include dideoxy-sequencing and virtually all genotyping platforms, especially those that depend on identification of specific polymorphisms, such as SNPs. Prior to PCR, molecular cloning was used as an approach to complexity reduction. Today, some form of PCR that specifically amplifies the locus or loci to be analyzed is most commonly employed. Typically, one PCR reaction is used to amplify a single locus of interest that is less than one thousand base pairs long. As a result, technologies aspiring to analyze many (hundreds to millions) loci are often limited by the cost of PCR amplification (~$0.10-$1.00 per reaction). One solution to the efficiency of PCR in complexity reduction is multiplex amplification, wherein multiple primer pairs specific for different loci are used to simultaneously amplify said loci in a single reaction. Practically speaking, few multiplex PCR strategies have resulted in the reliable amplification of more than 50 loci simultaneously. Thus, multiplex PCR represents a solution with limited scalability. Approaches that enable the isolation and amplification of thousands of loci in a single reaction are needed.

[0064] Detection of sample (target) hybridization to an oligonucleotide array requires that a minimum amount of the specific subset of the sample corresponding to a probe (cognate target) be physically associated with said probe. Hybridization of cognate target to a probe can be facilitated by: 1. increasing the total concentration of the sample (total target)

being exposed to the microarray, 2. increasing the time available for hybridization to occur, or 3. increasing the proportion of cognate target in the sample (total target). This third strategy also reduces the amount of non-specific hybridization of non-cognate target to a probe, thereby reducing background. This term is termed complexity reduction because it converts a DNA sample containing many loci (high complexity) to one containing fewer loci (low complexity).

[0065] One embodiment of the present invention involves the digestion of nucleic acids with a type IIs restriction enzyme to create fragments with overhangs of a desired length (e.g., 4 base pairs) and varied sequence, ligation of linkers that are specific to a subset of the overhangs produced by the restriction enzyme, and amplification by PCR with linker-specific primers.

[0066] Type IIs restriction enzymes are restriction enzymes that recognize sequences that are continuous and asymmetric (nonpalindromic sites) in a nucleic acid molecule. They bind to the nucleic acid as monomers and cleave both strands of the nucleic acid at a fixed distance outside of their recognition site through dimerization of the cleavage domains of adjacent enzyme molecules. Most of the type IIs restriction enzymes cleave outside of their recognition sequence to one side. although some type IIs restriction enzymes cleave outside of the recognition sequence on both sides, resulting in nucleic acid fragments that have variable sequences at one or both ends. Some examples of type IIs restriction enzymes are Fok I, Alw I, Bsg I, Bpm I and Mbo II. Further, the methods of the present invention may utilize a single type IIs restriction enzyme, or a combination of two or more type IIs restriction enzymes to fragment the original nucleic acid sample. The type IIs restriction enzymes described in the following examples cleave both strands outside of the recognition sequence resulting in a pool of nucleic acid fragments that contain variable sequence at both ends. In certain embodiments, and in the following examples, the type IIs restriction enzyme(s) used in the present invention cleaves the nucleic acid to produce single-stranded overhangs or "sticky ends" of variable sequence on the ends of the resulting nucleic acid fragments.

[0067] After cleavage with one or more type IIs restriction enzymes, linkers are annealed to the resulting nucleic acid fragments. Typically, the linkers are partially double-stranded, although fully single-stranded linkers may also be used. Each linker comprises a PCR primer binding site so that a "selected" subset of nucleic acid fragments (that which comprises those nucleic acid fragments that are bound by a linker at each end) may be PCR

amplified. Each linker also comprises a single-stranded region that binds to a single-stranded sticky end of a nucleic acid fragment. Since the single-stranded ends of the nucleic acid fragments comprise variable sequence, a single linker design comprises a single-stranded overhang that may be annealed to only the portion of the nucleic acid fragments to which it is complementary. Thus, some fragments are bound by a linker at both ends, while some are bound by a linker at only one end, and some are not bound by a linker at either end. Using different linkers comprising different overhangs that are complementary to the single-stranded ends of different nucleic acid fragment enables linker ligation to, and therefore PCR amplification of, different subsets of the nucleic acid fragments. Further, linkers with different overhangs may be used independently or in combination, thus "selecting" different subsets of the original pool of nucleic acid fragments based on the sequence of their single-stranded ends.

[0068] A single-stranded overhang of a linker may comprise a specific nucleotide sequence, in which case it would only bind to single-stranded fragment ends comprising the single complement to that specific sequence. For example, if a linker comprises an overhang of the sequence 5'-GCTT-3', then the only fragment ends to which it binds are those of the sequence 3'-CGAA-5'. Nucleic acid fragments that comprise this sequence on both ends can bind to the linker at both ends. Alternatively, a single-stranded overhang of a linker may comprise variable nucleotide positions to allow binding of the linker to a plurality of the fragments comprising all the complements of that pool of single-stranded overhangs. For example, a linker comprising an overhang of the sequence 5'-GCNT-3' (where "N" indicates a variable position and may be any nucleotide) is in actuality a pool of linkers, each of which comprise one of the following overhangs: 5'-GCAT-3', 5'-GCGT-3', 5'-GCCT-3', or 5'-GCTT-3'. The only fragment ends to which these linkers would bind are those of the following sequences, respectively: 3'-CGTA-5', 3'-CGCA-5', 3'-CGGA-5', and 3'-CGAA-5'. Nucleic acid fragments that comprise one of these sequences at each end bind linker at both ends. Therefore, increasing the number of variable nucleotide positions in linker overhangs increases the proportion of nucleic acid fragments to which the linker may be annealed, as does using a combination of linkers rather than just one. Thus, the ligation of linkers to the nucleic acid fragments is used to select or discriminate between different subsets of nucleic acid fragments.

[0069] For example, consider a pool of linkers, each comprising a four base pair overhang of the sequence "NNNT", wherein "N" can be any nucleotide and the fourth

position is "fixed" as a "T". Since only one nucleotide in the linker overhang is specified, about one out of every 16 (¼ x ¼) fragments is complementary to the linkers at both ends and therefore supports linker ligation to both ends of the fragment. In this example, the same criterion is met at both ends of the fragment: an "A" is present at the position in the nucleic acid fragment end that is complementary to the fixed position in the linker overhang. Amplicons resulting from PCR amplification of the fragments that are bound by linker at both ends therefore represents in a 16-fold reduction in the complexity of the original nucleic acid sample.

[0070] However, because the sequences of the single-stranded ends of the nucleic acid fragments are independent of one another, combinations of linkers are required to represent all fragments. For example, a combination of linkers (one with an overhang of 5'-NNNG-3' and the other with an overhang of 5'-NNNT-3') would be required to amplify fragments comprising 3'-NNNC-5' at one single-stranded end and 3'-NNNA-5' at the other single-stranded end. Of course, nucleic acid fragments comprising either 3'-NNNC-5' or 3'-NNNA-5' at both ends would also be bound by linker at both ends and hence also amplified in a subsequent PCR. It follows that in the original pool of nucleic acid fragments there are ten different combinations of ends that may be bound by linkers containing one fixed position: A+A, C+C, G+G, T+T, A+C, A+G, A+T, C+G, C+T, and G+T, where each nucleotide specified refers to the position in the single-stranded end that is complementary to the fixed position in the linker overhang. As such, six combinations of pairs of linkers comprising (A+C, A+G, A+T, C+G, C+T, G+T) are required to amplify all 10 different subsets of fragments. Clearly, some of the subsets of fragments will be amplified in more than one reaction comprising each of the combinations of linkers, as discussed above.

[0071] Further, the particular linker overhang sequences can be designed to enable amplification of larger (higher complexity) or smaller (lower complexity) subsets of the original pool of nucleic acid fragments. For example, the more specific the linker overhang sequence (the more fixed nucleotide positions in the sequence) the smaller the resulting subset, and the less specific the linker overhang sequence (the fewer fixed nucleotide positions in the sequence) the larger the resulting subset. In addition, knowledge of the nucleotide composition of the pool of nucleic acid fragments can be used to design linkers to select higher or lower complexity subsets of the pool.

[0072] In one embodiment of the present invention, all the linkers contain the same PCR primer binding site, and only one PCR primer is required to amplify all fragments that

are bound by a linker at each end. In this case, each ligation reaction must be separately carried out to select a particular subset of the original pool of nucleic acid fragments, and multiple similar ligation reactions must be performed if more than one subset of the pool of nucleic acid fragments is to be selected.

[0073] Such an example is shown in Figure 4A-B. In this example, a nucleic acid sample (400) is first digested with a type IIs restriction enzyme SfnN I, which recognizes the sequence of GATGC and cleaves upstream of this sequence (both strands) to create a single-stranded overhang of 4 base pairs (410). One region of this nucleic acid sample is zoomed in to show the enzyme recognition sequence (i.e. binding site) and the cleavage site (410). The horizontal arrows on top of the binding sites indicate the side outside of the binding site to which SfnN1 cleaves (400 and 410). The sequence of the single-stranded ends varies depending on the nucleic acid sequence at the cleavage site. Using one or a combination of two or more type IIs restriction enzymes (420), a pool of nucleic acid fragments that contain variable single-stranded sequences at one or both ends (430) can be generated. As shown in Figure 4B, linker with single-stranded overhang (440) is ligated onto the nucleic acid fragments that contain single-stranded end(s) having a sequence that is complementary to that of the linker overhang. In this example, all 4 nucleotides in the linker overhang are specified (e.g. AAAA), therefore it will only bind to nucleic acid fragments with single-stranded end sequence of TTTT (450). About one out of 65536 (¼ x ¼ x ¼ x ¼ x ¼ x ¼ x ¼ x ¼) fragments will have a single-stranded overhang sequence of TTTT at both ends, representing a 65536-fold reduction in complexity in this step alone. Finally, the complexity is reduced even further by using linkers with a specific PCR primer, thereby selectively amplifying the nucleic acid fragments with TTTT overhang at both ends (460).

[0074] In alternative embodiments, the linkers may be designed to comprise a different PCR primer binding site for each different overhang. This linker design allows a single ligation reaction to be performed with multiple different linkers in the same reaction vessel without compromising the ability to selectively amplify only a subset of the original pool of nucleic acid fragments since the selection or discrimination between subsets of nucleic acid fragments occurs during PCR rather than the ligation reaction, thereby requiring a single ligation reaction to anneal multiple linkers to the pool of nucleic acid fragments after which a subset of the nucleic acid fragments can be selectively amplified based on the PCR primers used in a subsequent PCR amplification. In these embodiments, the number of different PCR primer binding sites is equal to the number of different linker overhangs.

[0075] In certain embodiments, a first subset of linkers with different single-stranded overhangs, and hence different specificities for binding to the nucleic acid fragments, may contain the same PCR primer binding site, while another subset of linkers with single stranded overhangs that are distinct from the first subset may comprise a different PCR primer binding site than the first subset. In these embodiments, the number of different PCR primer binding sites is less than the number of different linker overhangs, therefore reducing the total number of different PCR primers required to amplify the entire pool of nucleic acid fragments.

[0076] One example of an embodiment illustrating selection of subsets of nucleic acid fragments by PCR is shown in Figure 4A-C. As previously described (in Figure 4A), a pool of nucleic acid fragments is generated from the original nucleic acid sample by using one or more type IIs restriction enzymes (or a combination of different types of restriction enzymes). In this method (Figure 4C), a pool of ligate linkers, each comprising a different single-stranded overhang sequence (represented by different patterns in 470), are used in the ligation reaction to bind to different nucleic acid fragments at the single-stranded ends that comprise the complement sequence (480). The linkers are designed to comprise a different PCR primer binding site for each different overhang sequence, which is represented by the different patterns used in the double-stranded linker sequence (470). For example, the PCR primer sequence specifically for one linker overhang is represented by using the same pattern in the double-stranded linker primer region as in the linker overhang (470). The complexity in the nucleic acid fragments is further reduced by selective amplification of different subsets of the nucleic acid fragments using different PCR primer pairs (490).

[0077] In another embodiment of the present invention, a PCR-based method for complexity reduction of a nucleic acid sample is provided that does not require a restriction enzyme digestion or a ligation reaction. This method comprises the following steps: denaturation of the nucleic acid sample, annealing of a double-stranded branch primer to the denatured nucleic acid sample, extension of the primer by a DNA polymerase, and a second round of denaturation, annealing and extension. The second round of denaturation, annealing and extension is followed by PCR, optionally conditional PCR in which only fragments of a specified size range will be amplified.

[0078] This method is of particular utility for amplifying a region between two instances of a given recognition sequence when the recognition sequence is inappropriate (e.g., too short) to serve as a PCR primer binding site. Standard PCR typically requires

primers that are about 20 nucleotides in length. This method allows one to amplify the region between recognition sequences that are much shorter while simultaneously providing longer recognition sequences for subsequent cycles of PCR utilizing standard PCR primers.

[0079] One embodiment of the present invention is shown in Figure 5. A double stranded nucleic acid (500) is denatured and annealed to a partially double-stranded primer (505) comprising a single-stranded tail with a nucleotide sequence that is complementary to a desired recognition sequence in the nucleic acid. The single-stranded tail of the primer binds to the recognition sequence in the nucleic acid (510) and serves as a primer for DNA polymerase, which will extend the primer along the length of the nucleic acid in the primer extension reaction (520). The denaturation, annealing, and primer extension by DNA polymerase is repeated at least once (525-535). From regions of the original nucleic acid molecule that comprise two recognition sequences that are close together, amplicons are generated that comprise the partially double-stranded primer at both ends (540). Subsequent PCR amplification (545) using single-stranded primers that are complementary to the partially double-stranded primer allows exponential amplification of the region between the recognition sequences. Optionally, "Conditional PCR" can be used to limit the size of amplicons so that only those recognition sequences within a certain number of base pairs of one another and the region specified in between will be amplified. For example, by changing some of the variables of the PCR reaction, such as temperature, salt concentration, and/or extension time, or any combination of these conditions, a selective subset of nucleic acid fragments that are 300 to 600 base pairs in size are amplified (550). Therefore, the method described in this embodiment not only allows selective amplification of a region between two instances of a given recognition sequences (i.e. selection of subsets of the original pool of nucleic acid fragments), it also allows further selection of these fragments based on the fragment size, thereby greatly reducing the complexity of the original nucleic acid sample.

[0080] In another embodiment of the present invention, a PCR-based method for complexity reduction of a nucleic acid sample is provided that uses a restriction enzyme that recognizes an interrupted palindrome to digest a nucleic acid sample, anneals an adaptor containing fixed bases onto the ends of the resulting nucleic acid fragments, and then amplifies the adaptor-bound nucleic acid fragments using primers that contain fixed bases that are complementary to only a subset of the nucleic acid fragments.

[0081] For example, Figure 6 shows one embodiment of the present invention in which the Hpy188 III restriction enzyme is used to digest a nucleic acid sample. The

recognition sequence of Hpy188 III is "TCNNGA" with cleavage occurring between the C and N (600). Since only one in 256 (¼ x ¼ x ¼ x ¼) fragments of the original nucleic acid sample will have the interrupted 4-base pair palindromic sequence TC/GA, which is required for the enzyme activity of Hpy188 III, using this restriction enzyme alone represents a 256-fold reduction in complexity (610). Alternatively, using other restriction enzymes that recognize an interrupted palindromic sequence with more base pairs or using a combination of different restriction enzymes, the complexity in a nucleic acid sample can be further reduced by selecting a smaller subset for analysis. For example, using restriction enzyme Bgl I, which recognizes an interrupted palindrome of GCCNNNN/NGGC, more than 4000-fold (4^6) reduction in complexity can be achieved. Following the restriction enzyme digestion, an adaptor is ligated onto the resulting nucleic acid fragments (615). The adaptor contains fixed nucleotides whereby it only anneals to a subset of the fragments (620). In this example, the adaptor contains "AA" so the only fragments that bind to the adaptor must have "TT" at one or both ends (615, 620). Only one in 256 (¼ x ¼ x ¼ x ¼) fragments will contain a "TT" at both ends, and so will incorporate the adaptor at both ends, representing a 256-fold reduction in complexity as compared to the original nucleic acid sample (625).

[0082] Next, PCR is carried out with primers that bind to the adaptor, so only fragments that contain an adaptor sequence at each end are amplified. Optionally, the PCR primers can extend past the recognition sequence of the restriction enzyme and incorporate additional nucleotides to further reduce the complexity of the resulting amplicons. In this example, the primers comprise a fixed nucleotide (T) such that extension can only occur if the ambiguous position adjacent to the restriction enzyme recognition sequence (the "-1" position) is an A (630). Therefore, the only fragments that are amplified are those that contain an adaptor sequence at both ends and those that contain an A in the -1 position adjacent to the restriction enzyme recognition sequence (635). The -1 fixed position provides another 16-fold reduction in complexity of the nucleic acid sample. Of course, as one of skill will readily recognize, additional fixed nucleotide positions may also be incorporated into the primer sequence to further reduce the complexity of the resulting pool of amplicons. Optionally, in addition to targeting the upstream sequence of the recognition site by using PCR primers with one or more fixed nucleotide positions, the complexity of the nucleic acid sample is further reduced by using "Conditional PCR" to limit the size of amplicons so that only those recognition sequences within a certain number of base pairs of one another and the region specified in between will be amplified. For example, a subset of nucleic acid fragments that

are 300 to 600 base pairs in size can be selectively amplified by changing the reaction condition, such as temperature, the salt concentration in PCR solution, and/or extension time, or any combination of these conditions.

Hybridization to Probe Arrays

[0083] Target nucleic acid samples prepared by one of the selection/enrichment methods described above are hybridized to microarrays for polymorphism detection. In one embodiment, more than one target sample can be hybridized simultaneously to a microarray and distinguished by the use of two-color labelling (e.g., the reference sequence bears one label and a target sample bears a second label). If the array is hybridized to a control reference sequence (or a target sequence that is identical to the reference sequence), all probes in the first probe set specifically hybridize to the reference sequence. If the array is hybridized to a target sample containing a target sequence that differs from the reference sequence at a polymorphic site, then probes flanking the polymorphic site do not show specific hybridization, whereas other probes in the first probe set distal to the polymorphic site do show specific hybridization.

[0084] The existence of a polymorphism also may be manifested by differences in normalized hybridization intensities of probes flanking the polymorphism relative to the probes when hybridized to corresponding targets from different individuals. For example, relative loss of hybridization intensity in a "footprint" of probes flanking a polymorphism signals a difference between the target and reference (i.e., a polymorphism) (see EP 717,113, incorporated by reference in its entirety for all purposes). Additionally, hybridization intensities for corresponding targets from different individuals can be classified into groups or clusters suggested by the data, not defined a priori, such that isolates in a given cluster tend to be similar and isolates in different clusters tend to be dissimilar. See WO 97/29212 (incorporated by reference in its entirety for all purposes).

[0085] Primary arrays of probes also can contain second, third and fourth probe sets as described in WO 95/11995. The probes from the three additional probe sets are identical to a corresponding probe from the first probe set except at the interrogation position, which occurs in the same position in each of the four corresponding probes from the four probe sets, and is occupied by a different nucleotide in the four probe sets. After hybridization of such an array to a labelled target sequence, analysis of the pattern of label should reveal the nature and position of differences between the target and reference sequence. For example,

comparison of the intensities of four corresponding probes reveals the identity of a corresponding nucleotide in the target sequences aligned with the interrogation position of the probes. The corresponding nucleotide is the complement of the nucleotide occupying the interrogation position of the probe showing the highest intensity.

[0086] Additionally, arrays for de novo polymorphism detection can tile both strands of reference sequences. Both strands are tiled separately using the same principles described above, and the hybridization patterns of the two tilings are analyzed separately. Typically, the hybridization patterns of the two strands indicate the same results (i.e., location and/or nature of polymorphic form) increasing confidence in the analysis. Occasionally, there may be an apparent inconsistency between the hybridization patterns of the two strands due to, for example, base-composition effects on hybridization intensities. Such inconsistency signals the desirability of rechecking a target sample either by the same means or by some other sequencing methods, such as use of an ABI sequencer.

[0087] Arrays used for analyzing previously identified polymorphisms typically differ from the arrays for de novo identification in the following respects. First, whereas probes are typically included to span the entire length of a reference sequence in de novo discovery arrays, in arrays for analyzing precharacterized polymorphisms, only a segment of a reference sequence containing a polymorphic site and immediately flanking bases typically is spanned. For example, this segment is often of a length commensurate with that of the probes. Second, an array for analyzing precharacterized polymorphisms typically includes at least two groups of probes. The first group of probes is designed based on the reference sequence, and the second group is designed based on a polymorphic form thereof. If there are three polymorphic forms at a given polymorphic site, a third group of probes can be included. Finally, because fewer probes are generally required to analyze precharacterized polymorphisms than in the de novo identification of polymorphisms, the former arrays often are designed to detect more different polymorphic sites than primary arrays. For example, whereas a de novo polymorphism discovery array may tile a single chromosome, an array for analyzing precharacterized polymorphisms can easily analyze 1,000, 10,000, 100,000 or 1,000,000 polymorphic sites in reference sequences dispersed throughout the human genome.

[0088] The design of suitable probe arrays for analysis of predetermined polymorphisms and interpretation of the hybridization patterns is described in detail in WO 95/11995; EP 717,113; and WO 97/29212. Such arrays typically contain first and second groups of probes, which are designed to be complementary to different allelic forms

of the polymorphism. Each group contains a first set of probes, which is subdivided into subsets, one subset for each polymorphism. Each subset contains probes that span a polymorphism and proximate bases and are complementary to one allelic form of the polymorphism. Thus, within the first and second probe groups there are corresponding subsets of probes for each polymorphism. The hybridization patterns of these probes to target samples can be analyzed by footprinting or cluster analysis, as described above. For example, if the first and second probe groups contain subsets of probes respectively complementary to first and second allelic forms of a polymorphic site spanned by the probes, then on hybridization of the array to a sample that is homozygous for the first allelic form, all probes in the subset from the first group show specific hybridization, whereas probes in the subset from the second group that span the polymorphism show only mismatch hybridization. The mismatch hybridization is manifested as a footprint of probe intensities in a plot of normalized probe intensity (i.e., target/reference intensity ratio) for the subset of probes in the second group. Conversely, if the target sample is homozygous for the second allelic form, a footprint is observed in the normalized hybridization intensities of probes in the subset from the first probe group. If the target sample is heterozygous for both allelic forms, then a footprint is seen in normalized probe intensities from subsets in both probe groups although the depression of intensity ratio within the footprint is less marked than in footprints observed with homozygous alleles.

[0089] Alternatively, the first and second groups of probes can contain first, second, third and fourth probe sets. Each of the probe sets can be subdivided into subsets, one for each polymorphism to be analyzed by the array. The first set of probes in each group spans a polymorphic site and proximate bases and is complementary to one allelic form of the site. The second, third and fourth sets, each have a corresponding probe for each probe in the first probe set, which is identical to a corresponding probe from the first probe set except at the interrogation position, which occurs in the same position in each of the four corresponding probes from the four probe sets and is occupied by a different nucleotide in the four probe sets.

[0090] Arrays for analyzing precharacterized polymorphisms are interpreted in similar manner to the arrays for polymorphism discovery having four sets of probes described above. For example, consider an array having first and second groups of probes, where each group has four sets of probes based on first and second allelic forms of a single polymorphic site. This array is then hybridized to a target containing a homozygous first allele. The

probes from the first probe set of the first group all show perfect hybridization to the target sample, and probes from the other probe sets in the first group all show mismatch hybridization. All probes from the second group of probes show at least one mismatch except the one of the four corresponding probes having an interrogation position aligned with the polymorphic site and having the same sequence as the first probe set of the first group that hybridized to the target. A probe from the second, third or fourth probe set having an interrogation position occupied by a base that is the complement of the corresponding base in the first allelic form shows specific hybridization.

[0091] If such an array is hybridized to a target sample containing a homozygous second allelic form, the mirror image hybridization pattern is observed. That is, all probes in the first probe set of the second group show matched hybridization, and probes from the second, third and fourth probe sets in the second probe group show mismatch hybridization. All but one probe in the first group of probes shows mismatch hybridization. The one probe showing perfect hybridization has an interrogation site aligned with the polymorphic site and occupied by the complement of the base occupying the polymorphic site in the second allelic form.

[0092] If such an array is hybridized to a target sample containing heterozygous first and second allelic forms, the aggregate of the above two hybridization patterns is observed. That is, all probes in the first probe set from both the first and second group show perfect hybridization (albeit with reduced intensity relative to a homozygous target), and one additional probe from the second, third or fourth probe set in each group shows perfect hybridization. In each group, this probe has an interrogation position aligned with the polymorphic site and occupied by a base occupying the polymorphic site in one or other of the allelic forms.

[0093] Typically, arrays for analyzing precharacterized polymorphisms contain multiple subsets of each of the probe sets described, with a separate subset for each polymorphism. Thus, for example, a secondary array for analyzing a thousand polymorphisms might contain first and second groups of probes, each containing four probe sets, with each of the four probe sets, being divided into 1000 subsets corresponding to the 1000 different polymorphisms. In this situation, analysis of the hybridization patterns from four subsets relating to any given polymorphism is independent of any other polymorphism. Analysis of the hybridization pattern of such an array to a target sample indicates which polymorphic form is present at some or all of the polymorphic sites represented on an array.

Thus, the individual is characterized with a polymorphic profile representing allelic variants present at a substantial collection of polymorphic sites.

[0094] Methods for using arrays of probes for monitoring expression of mRNA populations are described in WO 97/27317, and US 5.800.992. Some methods employ arrays having nucleic acid probes designed to be complementary to known mRNA sequences. mRNA populations or nucleic acids derived therefrom are applied to such an array, and targets of interest are identified. and optionally, quantified from the extent of specific binding to complementary probes. Optionally, binding of target to probes known to be mismatched with the target can be used as a measure of background nonspecific binding and subtracted from specific binding of target to complementary probes. Some methods employ arrays of random or arbitrary probes (also known as generic arrays). Such probes hybridize to complementary mRNA sequences present in a population, and are particularly useful for identifying and characterizing hitherto unknown mRNA species.

[0095] Arrays of probes immobilized on supports can be synthesized by various methods. Methods of forming arrays of nucleic acids. peptides and other polymer sequences are disclosed in. for example. 5.143,854. 5.252.743. 5.384.261, 5,405,783, 5,424,186, 5,445,943, 5,510.270. 5.677.195. 5.571,639. 6.040.138. all incorporated herein by reference for all purposes. The oligonucleotide array can be synthesized on a solid substrate by a variety of methods. including light-directed chemical coupling. and mechanically directed coupling. See US 5.143.854. and Fodor et al.. WO 92/10092 and WO 93/09668 and US 5,677,195, 6,040.193. and 5.831.070. Such arrays typically have at least 1000. 10,000, 100,000 or 1,000.000 different probes occupying 1000 different regions within a square centimeter. Algorithms for design of masks to reduce the number of synthesis cycles are described by Hubbel et al.. US 5.571,639 and US 5.593.839. Arrays also can be synthesized in a combinatorial fashion by delivering monomers to cells of a support by mechanically constrained flowpaths. or synthesized by spotting monomers reagents on to a support using an ink jet printer. by spotting preformed nucleic acid probes on to a substrate, or by covalently attached or attached via noncovalent linkage, such as biotin-avidin or biotin-streptavidin. Alternatively. the DNA can be held in place by coating the surface of an array with polylysine, which is positively charged and binds to negatively charged DNA. Nucleic acid probe arrays of standard or customized types are commercially available from Affymetrix, Inc. (Santa Clara. CA).

[0096] After hybridization of control and target samples to an array containing one or more probe sets as described above and optional washing to remove unbound and nonspecifically bound probe, the hybridization intensity for the respective samples is determined for each probe in the array. For fluorescent labels, hybridization intensity can be determined by, for example, a scanning confocal microscope in photon counting mode. Appropriate scanning devices are described by e.g., Trulson et al., US 5,578,832: Stern et al., US 5,631,734, and 60/223,278 filed 8/3/00, and 90/922,492, filed 8/3/01.

[0097] Reference sequences for polymorphic site identification are often obtained from computer databases such as Genbank, the Stanford Genome Center, The Institute for Genome Research, the Whitehead Institute, and the University of California at Santa Cruz. A reference sequence can vary in length from 5 bases to 100,000, 1 Mb, 10 Mb. 100 Mb or 1 GB bases. Reference sequences can be genomic DNA or episomes. In some methods, reference sequences are mRNA.

Applications

[0098] There are many applications for the methods of the present invention. For example, one can apply the methods of the present invention to association studies and diagnosis of disease. The polymorphic profile of an individual may contribute to phenotype of the individual in different ways. Some polymorphisms occur within a protein coding sequence and contribute to phenotype by affecting protein structure. The effect may be neutral, beneficial or detrimental, or both beneficial and detrimental. depending on the circumstances. For example, a heterozygous sickle cell mutation confers resistance to malaria, but a homozygous sickle cell mutation is usually lethal. Other polymorphisms occur in noncoding regions but may exert phenotypic effects indirectly via influence on replication, transcription, and translation. A single polymorphism may affect more than one phenotypic trait. Likewise, a single phenotypic trait may be affected by polymorphisms in different genes. Further, some polymorphisms predispose an individual to a distinct mutation that is causally related to a certain phenotype.

[0099] Phenotypic traits include diseases that have known but hitherto unmapped genetic components (e.g., agammaglobulimenia. diabetes insipidus, Lesch-Nyhan syndrome, muscular dystrophy. Wiskott-Aldrich syndrome, Fabry's disease, familial hypercholesterolemia, polycystic kidney disease, hereditary spherocytosis, von Willebrand's disease, tuberous sclerosis. hereditary hemorrhagic telangiectasia. familial colonic polyposis,

Ehlers-Danlos syndrome, osteogenesis imperfecta, and acute intermittent porphyria). Phenotypic traits also include symptoms of, or susceptibility to, multifactorial diseases of which a component is, or may be, genetic, such as autoimmune diseases, inflammation, cancer, diseases of the nervous system, and infection by pathogenic microorganisms. Some examples of autoimmune diseases include rheumatoid arthritis, multiple sclerosis, diabetes (insulin-dependent and non-independent), systemic lupus erythematosus and Graves disease. Some examples of cancers include cancers of the bladder, brain, breast, colon, esophagus, kidney, leukemia, liver, lung, oral cavity, ovary, pancreas, prostate, skin, stomach and uterus. Phenotypic traits also include characteristics such as longevity, appearance (e.g., baldness, obesity), strength, speed, endurance, fertility, and susceptibility or receptivity to particular drugs or therapeutic treatments.

[00100] Correlation is performed for a population of individuals who have been tested for the presence or absence of one or more phenotypic traits of interest and for polymorphic profile. The alleles of each polymorphism in the profile are then reviewed to determine whether the presence or absence of a particular allele is associated with the trait of interest. Correlation can be performed by standard statistical methods such as a chi-squared test and statistically significant correlations between polymorphic form(s) and phenotypic characteristics are noted. For example, it might be found that the presence of allele A1 at polymorphism A correlates with heart disease. As a further example, it might be found that the combined presence of allele A1 at polymorphism A and allele B1 at polymorphism B correlates with increased risk of cancer.

[00101] Such correlations can be exploited in several ways. In the case of a strong correlation between a set of one or more polymorphic forms and a disease for which treatment is available, detection of the polymorphic form set in a human or animal patient may justify immediate administration of treatment, or at least the institution of regular monitoring of the patient. Detection of a polymorphic form(s) correlated with serious disease in a couple contemplating a family may also be valuable to the couple in their reproductive decisions. For example, the female partner might elect to undergo in vitro fertilization to avoid the possibility of transmitting such a polymorphism from her husband to her offspring. In the case of a weaker, but still statistically significant correlation between a polymorphic set and human disease, immediate therapeutic intervention or monitoring may not be justified. Nevertheless, the patient can be motivated to begin simple life-style changes (e.g., diet, exercise) that can be accomplished at little cost to the patient but confer potential benefits in

reducing the risk of conditions to which the patient may have increased susceptibility by virtue of variant alleles. Identification of a polymorphic profile in a patient that correlates with enhanced receptiveness to one of several treatment regimes for a disease indicates that this treatment regime should be followed. For animals and plants, correlations between polymorphic profiles and phenotype are useful for breeding for desired characteristics.

[00102] Another application of the present invention is in the field of forensics. Determination of which polymorphic forms occupy a set of polymorphic sites in an individual identifies a set of polymorphic forms that distinguishes the individual. See generally, National Research Council, The Evaluation of Forensic DNA Evidence (Eds. Pollard et al., National Academy Press, DC, 1996). The more sites that are analyzed the lower the probability that the set of polymorphic forms in one individual is the same as that in an unrelated individual.

[00103] The capacity to identify a distinguishing or unique set of forensic markers in an individual is useful for forensic analysis. For example, one can determine whether a blood sample from a suspect matches a blood or other tissue sample from a crime scene by determining whether the set of polymorphic forms occupying selected polymorphic sites is the same in the suspect and the sample. If the set of polymorphic markers does not match between a suspect and a sample, it can be concluded (barring experimental error) that the suspect was not the source of the sample. If the set of markers does match, one can conclude that the DNA from the suspect is consistent with that found at the crime scene. If frequencies of the polymorphic forms at the loci tested have been determined (e.g., by analysis of a suitable population of individuals), one can perform a statistical analysis to determine the probability that a match of suspect and crime scene sample would occur by chance. If several polymorphic loci are tested, the cumulative probability of non-identity for random individuals becomes very high (e.g., one billion to one). Such probabilities can be taken into account together with other evidence in determining the guilt or innocence of the suspect.

[00104] An additional application of the methods of the present invention is the field of paternity testing. Paternity testing investigates whether the part of the child's genotype not attributable to the mother is consistent with that of the putative father. Paternity testing can be performed by analyzing sets of polymorphisms in the putative father and the child. If the set of polymorphisms in the child attributable to the father does not match the putative father, it can be concluded, barring experimental error, that the putative father is not the biological father. If the set of polymorphisms in the child attributable to the father does match the set of

polymorphisms of the putative father, a statistical calculation can be performed to determine the probability of coincidental match. If several polymorphic loci are included in the analysis, the cumulative probability of exclusion of a random male is very high. This probability can be taken into account in assessing the liability of a putative father whose polymorphic marker set matches the child's polymorphic marker set attributable to his/her father.

[00105] An additional important application of the present invention is in the field of expression analysis. The quantitative monitoring of expression levels for large numbers of genes can prove valuable in elucidating gene function, exploring the causes and mechanisms of disease, and for the discovery of potential therapeutic and diagnostic targets. Expression monitoring can be used to monitor the expression (transcription) levels of nucleic acids whose expression is altered in a disease state. For example, a cancer can be characterized by the overexpression of a particular marker such as the HER2 (c-erbB-2/neu) protooncogene in the case of breast cancer.

[00106] Expression monitoring can be used to monitor expression of various genes in response to defined stimuli, such as a drug. This is especially useful in drug research if the end point description is a complex one: i.e., not simply asking if one particular gene is overexpressed or underexpressed. Therefore, when a disease state or the mode of action of a drug is not well characterized, the expression monitoring can allow rapid determination of the particularly relevant genes.

[00107] In arrays of random probes (sometimes known as generic arrays), the hybridization pattern is also a measure of the presence and abundance of relative mRNAs in a sample, though it is not immediately known which probes correspond to which mRNAs in the sample. However the lack of knowledge regarding the particular genes does not prevent identification of useful therapeutics. For example, if the hybridization pattern on a particular generic array for a healthy cell is known and is significantly different from the pattern for a diseased cell, then libraries of compounds can be screened for those that cause the pattern for a diseased cell to become like that for the healthy cell. This provides a detailed measure of the cellular response to a drug.

[00108] Generic arrays also can provide a powerful tool for gene discovery and for elucidating mechanisms underlying complex cellular responses to various stimuli. For example, generic arrays can be used for expression fingerprinting. Suppose it is found that the mRNA from a certain cell type displays a distinct overall hybridization pattern that is

different under different conditions (e.g., when harboring mutations in particular genes, in a disease state). Then this pattern of expression (an expression fingerprint), if reproducible and clearly differentiable in the different cases can be used as a diagnostic. It is not required that the pattern be fully interpretable, but just that it is specific for a particular cell state (and preferably of diagnostic and/or prognostic relevance).

[00109] Both customized and generic arrays can be used in drug safety studies. For example, if one is making a new antibiotic, then it should not affect significantly the expression profile for mammalian cells. The hybridization pattern can be used as a detailed measure of the effect of a drug on cells, for example, as a toxicological screen.

[00110] The sequence information provided by the hybridization pattern of a generic array can be used to identify genes encoding mRNAs hybridized to an array. Such methods can be performed using DNA tags of the invention as the target nucleic acids described in WO 97/27317. DNA tags can be denatured forming first and second tag strands. The denatured first and second tag strands are then hybridized to the complementary regions of the probes, using standard conditions described in WO 97/27317. The hybridization pattern indicates which probes are complementary to tag strands in the sample. Comparison of the hybridization pattern of the two samples indicates which probes hybridize to tag strands that derive from mRNAs that are differentially expressed between the two samples. These probes are of particular interest, because they contain complementary sequence to mRNA species subject to differential expression. The sequence of such probes is known and can be compared with sequences in databases to determine the identity of the full-length mRNAs subject to differential expression provided that such mRNAs have previously been sequenced. Alternatively, the sequences of probes can be used to design hybridization probes or primers for cloning the differentially expressed mRNAs. The differentially expressed mRNAs are typically cloned from the sample in which the mRNA of interest was expressed at the highest level. In some methods, database comparisons or cloning is facilitated by provision of additional sequence information beyond that inferable from probe sequence by template dependent extension as described above.

EXAMPLES

Example 1: Isolation of cytoplasmic RNA from tissue culture cells:

[00111] In addition to using the methods of the present invention with cloned or genomic DNA, RNA may be used as a nucleic acid source for analysis. To prepare cytoplasmic RNA, cells were washed by adding 1 ml ice-cold PBS to a 10 cm tissue culture dish, and detaching the cells with a cell scraper. The cells were transferred to a 1.5 ml Eppendorf tube and centrifuged at 3000 rpm for 30 seconds. The supernatant was discarded and the cells were then suspended in 375 µl ice-cold lysis buffer (50mM Tris-Cl, pH 8.0: 100mM NaCl; 5mM $MgCl_2$, and 0.5% (v/v) nonidet P-40) and incubated on ice for 5 minutes. The samples were then centrifuged, and the supernatants were removed and placed in clean tubes containing 8 µl 10 % SDS. 2.5 µl of 20 mg/ml Proteinase K was then added to each tube and the samples were incubated at 37 ° C for 15 minutes. 400 µl of phenol/chloroform/isoamyl alcohol was then added, the tubes were shaken, then centrifuged for 10 minutes at room temperature. The aqueous phase was removed, and the extraction was repeated. An additional extraction was done with 400 µl chloroform. Again, the aqueous layer was removed and the RNA was precipitated with 1 ml 100% ethanol and 40µl 3M sodium acetate at pH 5.2. After precipitation, the pellets were rinsed with 1 ml 75% ethanol and 25% 0.1M sodium acetate, pH 5.2. Finally, the pellets were air dried and resuspended in 100 µl DEPC treated water. First strand cDNA synthesis was then carried out using the Life Technologies SuperScript II First Strand Synthesis kit (Life Technologies, Inc., Gaithersburg, MD).

Example 2: Second strand cDNA synthesis and adapter ligation

[00112] Once RNA was isolated, cDNA was prepared to be used in the methods of the present invention. First, 4 µl 10x buffer (500mM Tris-HCl pH 7.8, 50 mM $MgCl_2$, 100 µg BSA), 8 µl 0.4 mM dNTP, 20 µl first strand synthesis product, 2 µl DNA polymerase I (20U/µl), 2 µl RNase H (4U/µl), and water were combined and incubated at room temperature for one hour. Next, 10 µl 5x buffer, 0.25 µl DTT (100mM) and 2 µl T4 DNA polymerase (10U/µl) were added to the samples and incubated at 11 °C for 30 minutes. One volume of phenol-chloroform was then added, the tubes were centrifuged, and the upper layer was extracted with an equal volume of chloroform. The DNA was precipitated with 12.5 µl NaOAc (3M), 200 µl EtOH (100 %), and 12.5 µl glycogen (500 µg/ml) and overnight incubation at –20 °C. The DNA was then pelleted by centrifuging for 1 hour at 4 °C, the pellet was washed with 500 µl of 70% ethanol, and resuspended in 23 µl of water.

[00113] The double-stranded, blunt-ended DNA products were then ligated to adapters by adding 2 µg of the DNA to 3 µl adapters (1 µg/µl). 3 µl 10x T4 DNA ligase buffer and T4 DNA ligase (400U/µl) and incubating at room temperature overnight. The DNA products were purified through a Sephadex G-50 column and ethanol precipitated. Pellets were resuspended in buffer.

Example 3: Biotin labeling of DNA

[00114] Biotinylated residues were incorporated into target DNA using nick translation. The target DNA was DNA prepared by PCR amplification or a previously cloned DNA fragment, and other preparations known to those skilled in the art. The reactions were prepared by combining 1 µl purified DNA (0.1 mg/ml), 1 µl biotin 16-dUTP (0.04 mM), 2 µl 10x nick translation buffer (500 mM Tris-HCl (pH 7.5), 100 mM $MgCl_2$, 50 mM DTT), 1 µl dNTP mix (0.4 mM), [α-^{32}P]dCTP (3000 Ci/mmole), 1 µl DNAse I (10 mU), and water to 20 µl. The reaction mixture was incubated at 16 °C for 2 hours, then purified by spin column chromatography through Sephadex G-50 and ethanol precipitation. The pellet was resuspended in 10 µl buffer.

Example 4: Direct cDNA selection (primary selection)

[00115] Repeat sequences in the cDNA were blocked. This was performed by combining 5 µl of human genomic C_ot1 DNA (1µg) with 5 µl of the linker-adapted cDNA (1 µg). The reaction mixture was overlayed with mineral oil and heated for 10 minutes at 100 °C. The reaction was cooled to 65 °C and 10 µl of 2x hybridization solution (1.5 M NaCl, 40 mM Na phosphate buffer (pH 7.2), 10 mM EDTA (pH 8.0), 10x Denhardts solution, 0.2% SDS) was added to the reaction mixture under the oil. This mixture was then incubated for 4 hours at 65 °C. After hybridization, 5 µl of biotinylated (50 ng) target DNA was denatured and combined with 20 µl of the blocked DNA and 5µl of 2x hybridization solution (1.5 M NaCl, 40 mM Na phosphate buffer (pH 7.2), 10 mM EDTA (pH 8.0), 10x Denhardts solution, 0.2% SDS). This reaction was incubated for 2 days at 65 °C.

Example 5: Strepavidin–coated paramagnetic bead preparation

[00116] 3 mg of beads were washed three times with 300 µl of strepavidin bead-binding buffer (10 mM Tris-HCl (pH 7,5), 1 mM EDTA (pH 8.0), 1M NaCl) and the beads

were resuspended in a final concentration of 10 mg/ml in the buffer. An aliquot of each labeling reaction was tested for the ability to bind the beads by combining 20 µl of the beads with 1 µl labeled DNA (10 ng/µl) and 29 µl bead binding buffer and incubating at room temperature for 15 minutes. The beads were removed by using a magnetic separator and transferred to a fresh tube. The radioactivity was then measured and the binding considered successful if the ratio of bound to free cpm was >8:1.

Example 6: Binding of selected cDNA to strepavidin-coated paramagnetic beads

[00117] The DNA was then captured by combining 50 µl strepavidin-coated beads, 30 µl of the annealed reaction mix and 50 µl strepavidin bead-binding buffer (10 mM Tris-HCl (pH 7.5), 1 mM EDTA (pH 8.0), 1M NaCl). The mixture was incubated for 15 minutes at room temperature. The beads were removed using a magnetic separator and the supernatant was discarded. The beads were washed twice in 1 ml of 1 x SSC/0.1% SDS at room temperature followed by three washes, 15 minutes each in 1 ml 0.1x SSC/0.1%SDS at 65 °C. After the final wash, the beads were transferred to a fresh tube. Hybridized DNAs were eluted by adding 100 µl of 0.1M NaOH and incubating the reaction mixture for 10 minutes at room temperature. The mixture was desalted by spin-column chromatography through Sephadex G-50.

Example 7: Amplification of selected DNAs

[00118] Three aliquots (1 µl, 5 µl and 10 µl) of eluted cDNA were combined with 5 µl primer (10mM), 2.5 µl 10x amplification buffer, 2.5 µl dNTP mixture for PCR (2.5 mM), 0.2 µl Taq polymerase (5U/µl) and water to bring the final volume to 25 µl. In addition, control reactions were set up. The negative control did not have the eluted DNA added, and the positive control added sample DNA that had not gone through the biotin labeling and selection steps. DNA was amplified using 30 cycles of denaturation at 94 °C for 30 seconds, annealing at 55 °C for 30 seconds and polymerization at 72 °C for 1 minute. Aliquots of the reaction products (0.5 µg/lane) were loaded onto a 1% agarose gel. Once the enrichment was confirmed, the amplification reaction was scaled up to yield at least 1.5 µg of selected DNAs. The pooled reactions were extracted with phenol:chloroform and the DNA was recovered by ethanol precipitation. The DNA was air dried and resuspended in buffer.

[00119] Secondary selection was carried out under the same conditions as the primary selection using 1 µg of selected DNA and 50 ng of target DNA. Repetitive sequences were blocked with 1 µg of the selected DNA being used in the reaction. The final amplification products were visualized on an agarose gel.

Example 8: Preparing target DNA for hybridization

[00120] After reducing sample complexity (and optionally labeling) target DNA was prepared for application to a chip as follows: 177 µl 5M TMACL, 3 µl 1M Tris (pH 7.8 or 8), 3µl 1% triiton X-100, 3µl 10 mg/ml herring sperm DNA, 3µl 5nM control oligo, and labeled DNA and H_2O to achieve a 300 µl final volume. In various embodiments, the concentration of labeled DNA ranged from about 0.1pM to 100pM. The samples were denatured at 99°C for 5 minutes and spun down. The nucleic acid arrays were warmed to 50°C about 20 minutes before adding the hybridization mixture. The sample nucleic acids were then added to a chamber containing the array, hybridized at 50°C in a rotisserie using a rotation speed of 40 rpm.

Example 9: Staining and scanning an array

[00121] This example illustrates a procedure for detecting hybridization of sample to probes on an array.

Solutions:

1. Streptavidin-phycoerythrin Solution

   1ml total (300µl/chip)

   470µl water

   500µl 2X MES

   20µl acetylated BSA(50 mg/ml)

   10µl streptavidin-phycoerythrin(1mg/ml)

2. Antibody solution

   1ml total (300ul/chip)

   470µl water

   500µl 2X MES

   20µl acetylated BSA(50mg/ml)

   10µl biotinylated anti-streptavidin(1mg/ml)

Procedures:

[00122] First, a fluidics station (available from Affymetrix, Inc., Santa Clara) was primed with 6xSSPE/0.01% Triton X-100, and a scanner (also available from Affymetrix) was activated and an experimental information file was prepared according to the manufacturer's instructions. Hybridization solution was removed from the array and stored at -20°C. The array was then rinsed twice with 1x MES/0.01% Triton X-100, 300µl streptavidin solution was added, and the arrays were incubated at room temperature for 20 minutes. The stain solution was then removed and the array was rinsed twice with 1x MES/0.01% Triton X-100. Next, 300µl antibody solution was then added to the array and incubated at room temperature for 20 min. The antibody solution was removed and the array was rinsed twice with 1X MES/0.01% Triton X-100. 300µl staining solution was again added to the array and incubated at room temperature for 20 min. The array was then inserted into the fluidics station and washed 6 times at 35°C with 6X SSPE/0.01%Triton X-100. The array was then scanned.

Example 10: Fragmentation and labeling of genomic DNA or PCR fragments

[00123] To fragment and label genomic DNA, the following reagents were combined: 30 ul of purified DNA sample (400 ng) and 3.7 ul of 10x buffer. Just before placing the sample into 37°C water bath, 1 ul of 0.07U DNaseI was added into the sample mixture (DNaseI dilution: 1.4 ul of DNaseI + 18.6 ul cold 10 mM Tris, pH 8.0. Final concentration is 0.07U/ul). The samples were mixed and incubated at 37°C for 7 minutes. Next, the samples were heated at 99°C for 10 min to inactivate the DnaseI, and then cooled on ice for 2 minutes. The samples were centrifuged at a maximum speed of 14,000 rpm for 20 seconds.

[00124] To label the fragmented DNA, 1 ul of TdT and 1 ul of biotin-ddATP were added to the fragmented DNA sample. The samples were mixed and centrifuged at a maximum of 14,000 rpm for 20 seconds. The samples were then incubated at 37°C for 90 minutes and then at 99°C for 10 minutes to inactivate the TdT enzyme. The samples were then cooled on ice for 2 minutes, centrifuged, and kept on ice until ready for hybridization.

[00125] An alternative procedure for fragmenting by DNaseI digestion and labeling that is particularly suitable for use with long range PCR products uses long range PCR products in a volume of 300-350µl were obtained. The concentration of DNA was

determined by OD_{260} measurement. Next, 280 µg DNA was labelled to give a final target concentration of 5-10pM for a complexity range of 3-6 MB. The labeling was performed in five independent Eppenderf tubes with each one containing 37µl 10X One-Phor-All Buffer PLUS, 2 µl Gibco DNaseI (at 0.5U/uL), 1 µl Dnase 1, purified LR-PCR products up to 330µl in volume for a total reaction volume of 370µl, each tube was incubated at 37°C for 10 minutes, 99°C for 10 minutes, and 25°C for <5 minutes, and then spun briefly. 20 µl TdT (25 U/µl) and 20 µL biotin ddATP (1 mM) were then added to each tube, and then the tubes were incubated at 37°C for 90 minutes, 99°C for 10 minutes and 25°C for <5 minutes.

Example 11: Removal of repeat sequences

[00126] In an alternative protocol to remove repeat sequences, human placenta DNA was digested with DNaseI as follows: 160µg human placenta DNA (0.08fM for the full length) was added to 220µl reaction solution (64µl DNA (2.5ug/µl). 22µl 10X buffer, 3.5µl DNaseI (0.35U), 132µL wafer). 9 µl of 480mM $NaPO_4$ buffer, pH 7.4 was then added to reach a final $NaPO_4$ concentration of 126mM and a volume of 301µl. The sample was denatured for 5 minutes at 99°C, incubated at 65°C for 90 minutes to allow repeat sequences to hybridize, then diluted to 10mM $NaPO_4$ for HPLC.

Example 12: HPLC hydroxyapatite chromatography

[00127] This protocol illustrates use of a hydroxyapatite column to separate single-stranded and double-stranded DNA. One application of this protocol used single-stranded fragments with an average length 60 bases from chromosome 21 and double-stranded fragments of herring sperm DNA (average length 500 bp). Both single- and double-stranded DNA were present at 9µM. The column was an Econo-Pac CHT-II Cartridge having a DNA capacity of 160µg. The column was loaded with DNA in 10mM phosphate. At 10-20 mM phosphate hydroxyapatite binds both single and double stranded DNA. DNA was then eluted at a gradient from 10 mM to 1 M $NaPO_4$ buffer, pH 7.4 over 30 min. Elution was monitored by absorbance at 260 nm. At 5 minutes, there was a small peak indicating release of single stranded DNA, and at 25 minutes there was a larger peak indicating release of double stranded DNA, as shown in Fig. 1.

[00128] Additional methodology useful for practicing the invention are described in Birren et al. supra. All publications and patent applications cited above are incorporated by

reference in their entirety for all purposes to the same extent as if each individual publication or patent application were specifically and individually indicated to be so incorporated by reference. Although the present invention has been described in some detail by way of illustration and example for purposes of clarity and understanding, it will be apparent that certain changes and modifications may be practiced within the scope of the appended claims.